FasterAnalytics. An efficient implementation of Bayesian Networks for Modeling Large Datasets

Jorge Moraleda

Abstract

This white paper provides a brief introduction to the technology behind the FasterAnalytics software. FasterAnalytics' core technology is based on the Bayesian Networks paradigm. Bayesian Networks, also known as belief networks, are graphical models that can encode the probabilistic relationships among attributes in a certain domain.

Bayesian Networks have a number of properties; transparency, ability to deal with uncertainty and intuitive usage that make them attractive for a large number of analytic modeling applications.

FasterAnalytics has algorithmic enhancements that enable it to learn Bayesian Networks efficiently from large datasets. FasterAnalytics allows the user to interact easily with the models learned as well as make predictions about future data.

1 Bayesian Networks

Probabilistic graphical models are graphs in which nodes represent random variables, and the edges (lack of) represent conditional independence assumptions. In essence, they provide a compact representation of joint probability distributions for a given domain.

Bayesian Networks are a type of probabilistic graphical models with directed edges. Directed edges impose an order on the pair of nodes that they link. The edges go from a parent node to a child node [3, 7].

To fully capture this joint probability distribution, in addition to the graph structure it is necessary to specify the parameters of the model. For a Bayesian Network, one must specify the Conditional Probability Distribution (CPD) at each node. If the variables are discrete, this can be represented as a Conditional Probability Table (CPT), which lists the probability that the child node takes on each of its different values for each combination of values of its parents. Consider the following very small example for a lung disease diagnostic, in which all nodes are binary (i.e., have two possible states denoted by T (true) and F (false)).

There are two possible variables which, if known, would affect the probability of the event "person has dyspnoea–shortness of breadth" (D=true): lung cancer or bronchitis. In this model we can use our domain knowledge to choose the term "causes" of Dyspnoea to refer to Lung Cancer and Bronchitis. (Note that strong correlation does not always imply a causal relationship.) In this model there is a very small chance (1/1000) that the person will have shortness of breath without having either of the two parent conditions. This takes into account non-modelled causes for dyspnoea.

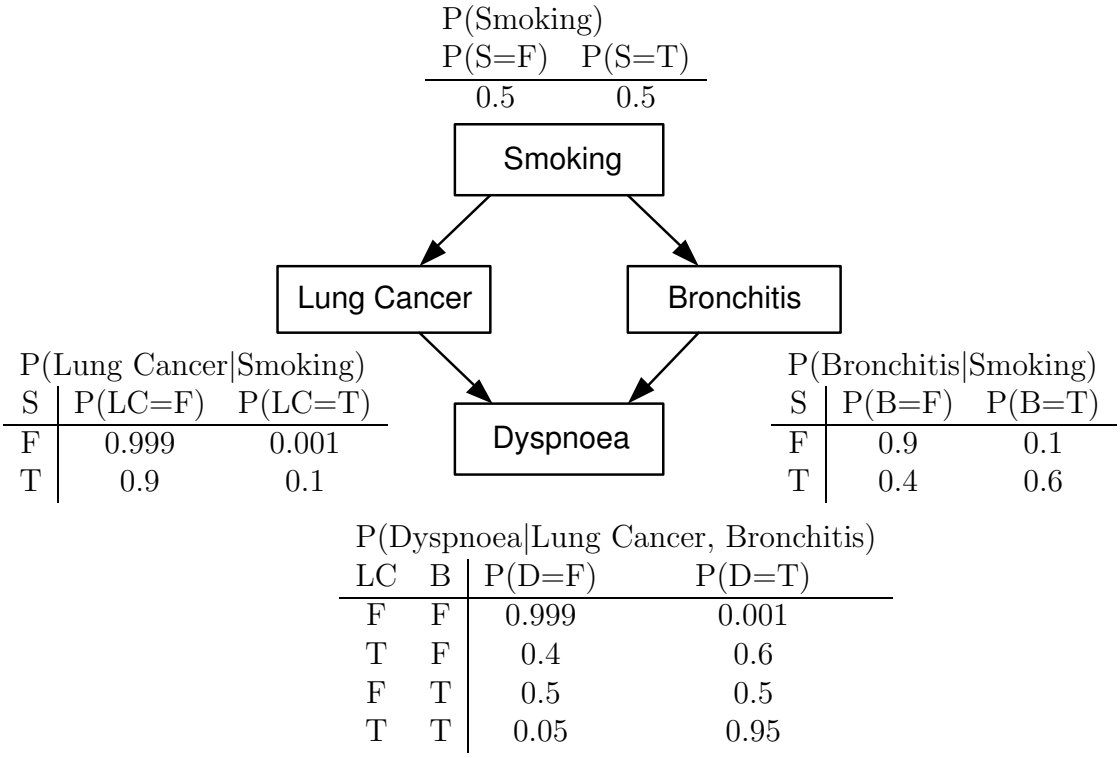

Figure 1: Small Bayesian Network for lung disease diagnostic.

The strength of the dependence relationships is shown in the table. For example $P(D = T \mid LC = T, B = F) = 0.6$ (the probability of Dyspnoea given Lung Cancer and not Bronchitis), and hence, $P(D = F \mid LC = T, B = F) = 1 - 0.6 = 0.4$, since each row must sum to one. Since the S node has no parents, its CPT specifies the prior probability that a person is a smoker (in this case, 0.5).

The conditional independence relationship encoded in a Bayesian Network can be stated as follows: a node is independent of its non-descendants given its parents, where the descendant relationship is defined as the transitive closure of the child relationship.

By the chain rule of probability, the joint probability of all the nodes in the graph above is: $P(S, LC, B, D) = P(S) * P(LC \mid S) * P(B \mid LC, S) * P(D \mid B, LC, S)$

However this representation can ge simplified using the conditional independence information in the Bayesian Network: D is conditionally independent of S if B and LC are known, likewise B is conditionally independent of LC given S. Thus, one can rewrite the joint probability distribution as $P(S, LC, B, D) = P(S) * P(LC \mid S) * P(B \mid S) * P(D \mid B, LC)$

It can be seen that the conditional independence relationships allow us to represent the joint probability distribution more compactly. Here the savings are minimal, but in general, if there were n binary nodes, the full joint distribution would require $O(2^n)$ space to represent, but the factored form would require $O(n2^k)$ space to represent, where k is the maximum number of parents of a node. Fewer parameters makes learning easier, as less data is required to determine the value of the parameters without overfitting.

2 Creation: Learning

Traditionally Bayesian Networks were acquired from experts in relevant fields through elicitation. This process is difficult because human beings often have difficulty expressing their knowledge in a quantitative form [2], and also expensive as experts' time is highly valuable.

More recently, many groups have investigated various techniques for learning Bayesian Networks from data.

Learning from data is a hard problem. In particular learning Bayesian Network structure from data is NP-hard [1, 9]. Thus heuristic search is necessary to find good models. There are two approaches to improve heuristic search: Speeding up model evaluation in order to be able to search through a larger number of models in a given search time and using better heuristics to generate higher quality models early in the search. FasterAnalytics uses both approaches by using two complementary advanced algorithms: AD+Tree [6] and Queue Learning [5].

The AD+Tree is a data structure that caches counts from the dataset very efficiently, enabling very fast evaluation of larger models. Under certain assumptions the AD+Tree allows one to process datasets one order of magnitude larger than those processed with other data structures of similar speed performance.

Queue Learning is an algorithm for learning Bayesian Network structure that has been shown to produce better models early in the search than existing techniques when applied to large datasets. Queue learning is also an inherently parallel algorithm thus holding the potential for significant speed improvements when used in distributed systems.

3 Usage: Insights and Inference

The two most common goals of learning Bayesian Networks from data are: gaining insights about the domain, and probabilistic inference (making predictions).

Insights can be gained because Bayesian Network structure generally reflects the underlying structure of the domain. In particular, users find it easy to interpret Bayesian Networks, because often the relationships have a causal interpretation [10, 8]. In order to gain insights it is important to be able to present the information contained in the Bayesian Network, both structural and quantitative, in an efficient manner. This is especially challenging in large complex networks. Existing user interfaces do not scale well with model size. FasterAnalytics uses new user interfaces that address the new challenges that that displaying and interacting with these larger models pose. In particular the usage of coloring schemes has increased the number of attributes that can be manipulated comfortably from a few tens to several hundreds.

The other goal of learning Bayesian Networks from data is making predictions about unobserved data. This is known as probabilistic inference. That is, given a set of observations of the values of some of the attributes in the domain, find the conditional probability distributions of one or more of the remaining attributes. That is the Bayesian Network allows one to predict the likelihood of the various states in non observed variables based on the states of observed variables. FasterAnalytics uses efficient exact inference algorithms [4] .

References

[1] D. M. Chickering. *Learning Bayesian Networks is NP-complete*, pages 121–130. Stringer Verlag, New York, 1996.

[2] M. Druzdzel, L. Van der Gaag, M. Henrion, and F. Jensen. Building probabilistic networks: Where do the numbers come from. In *International Joint Conferences in Artificial Intelligence (IJCAI) Workshop*, Dordrecht, The Netherlands, 1995.

[3] R. Howard and J. Matheson, editors. *Readings on the Principles and Applications of Decision Analysis*, volume 2, pages 721–762. Strategic Decisions Group, Menlo Park, CA, 1996.

[4] F. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, New York, 1996.

[5] J. Moraleda. Ad+tree: A compact adaptation of dynamic ad-trees for efficient machine learning on large data setsnew algorithms, data structures, and user interfaces for machine learning of large datasets with applications. *Ph.D. dissertation*, 2004.

[6] J. Moraleda and T. Miller. Ad+tree: A compact adaptation of dynamic ad-trees for efficient machine learning on large data sets. *Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning*, 2002.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffmann, San Mateo, CA, 1988.

[8] J. Pearl. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

[9] R. W. Robinson. *Counting Unlabeled Acyclic digraphs*. Springer-Verlag, New York, 1977.

[10] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.